Module II-b
Data Understanding

아이디케이스퀘어드 양승준 / sidney.yang@idk2.co.kr
https://www.heartcount.io

Data Description (Data Profiling)

데이터의 특성과 모양을 요약하여 기술하는 방법

Central Tendency
중심 경향

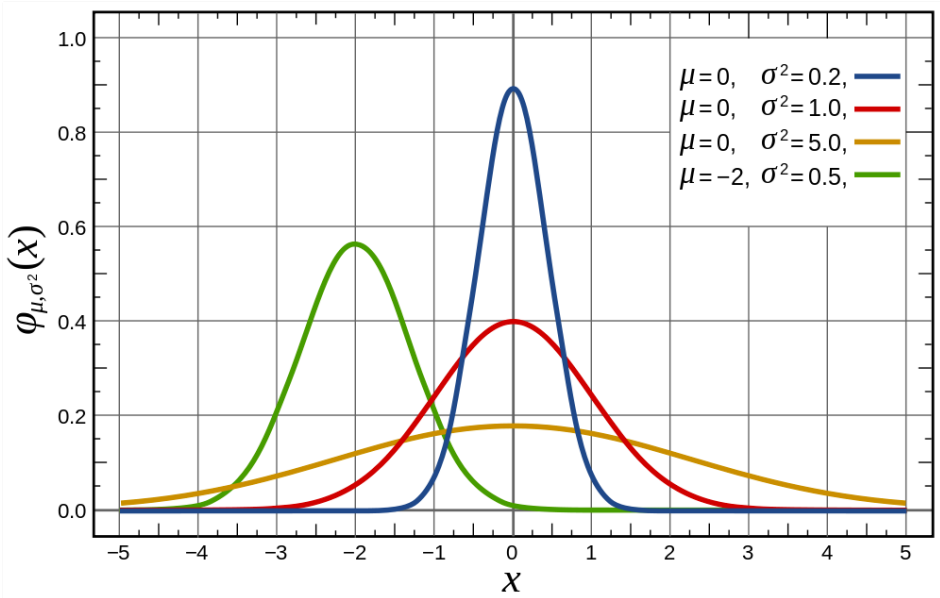
- 평균(Mean)
- 중앙값(Median)
- 최빈값(Mode)

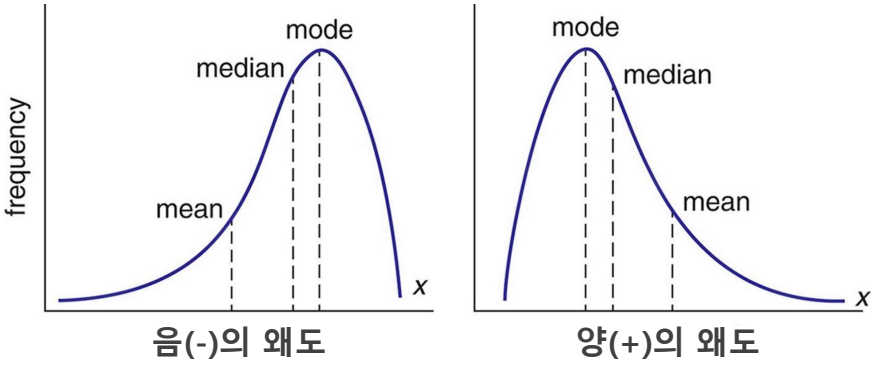
Dispersion
퍼진 정도

- 범위(Range)
- 분산(Variance)
- 표준편차(SD)
- Percentile

Shape of Distribution
퍼진 모양(대칭)

- 왜도 (Skewness)



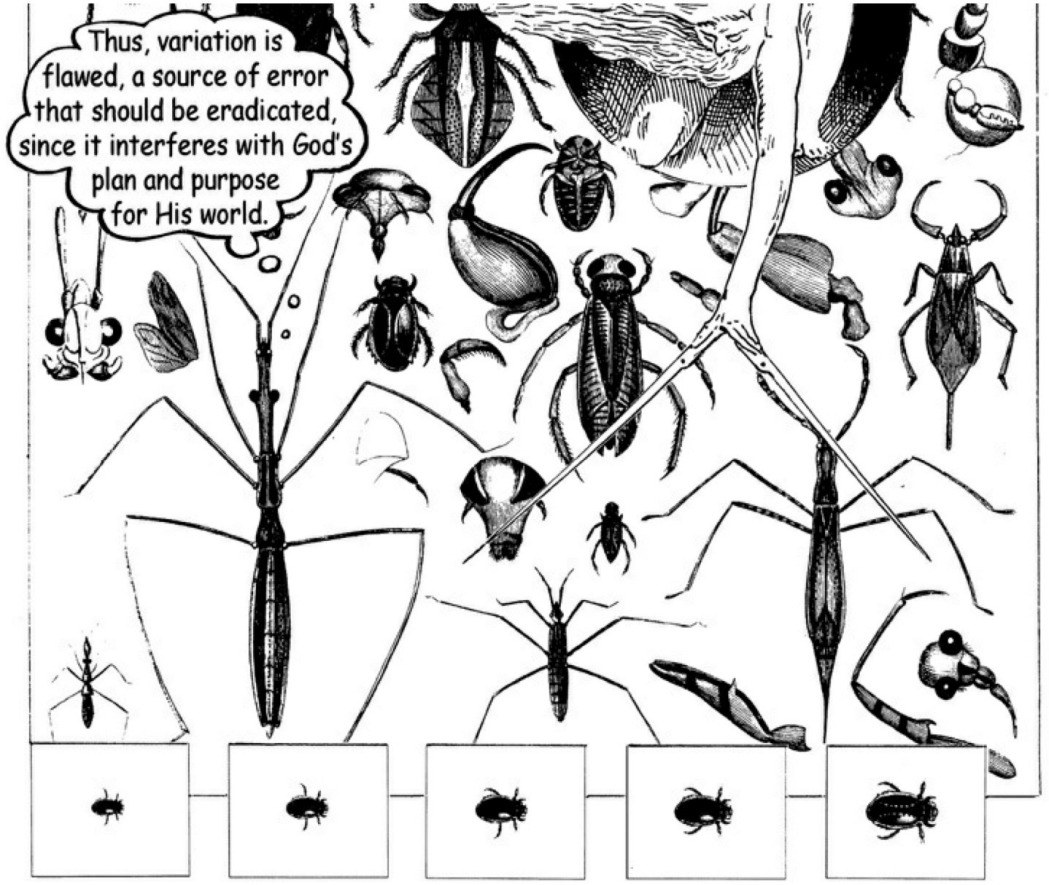


The Philosophy of Statistics [19th Century]

초기의 통계학 - 결정론적 세계관에 바탕을 둔 이데아/본질의 추구

평균값이 대상이 보유한 이상적인 속성이고(Idealized Mean)
Variation(차이)은 제거해야 할 오류라는 생각이 지배적이었음

Species(종; 種): An Ideal Type

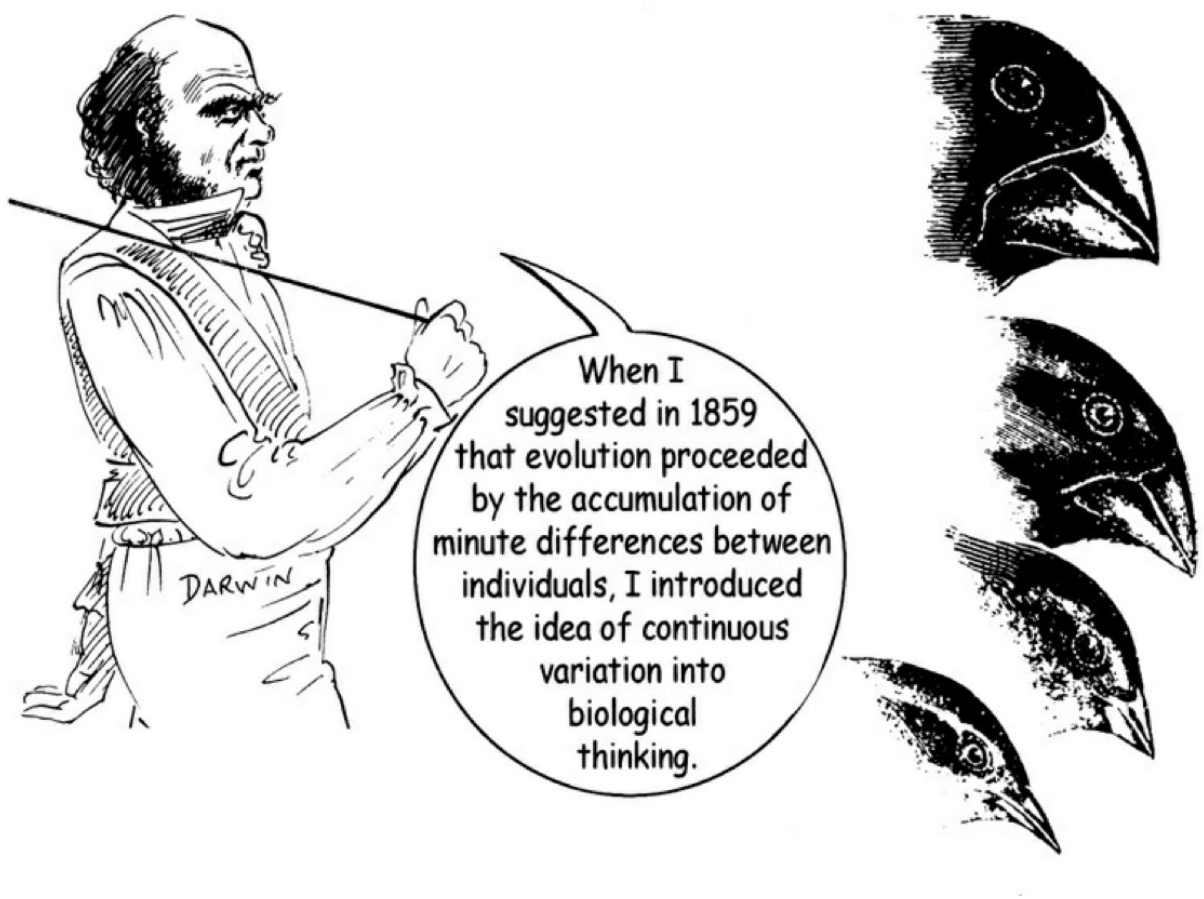


Darwin and Statistical Population [Late 19th~Early 20th Century]

다윈의 등장: Type/Essence(본질) → Variation(차이)

차이(변이)의 점진적 누적에 의해 진화가 이루어진다는 발견

개별 개체에 존재하는 의미있는 차이(변이)에 관심을 갖기 시작



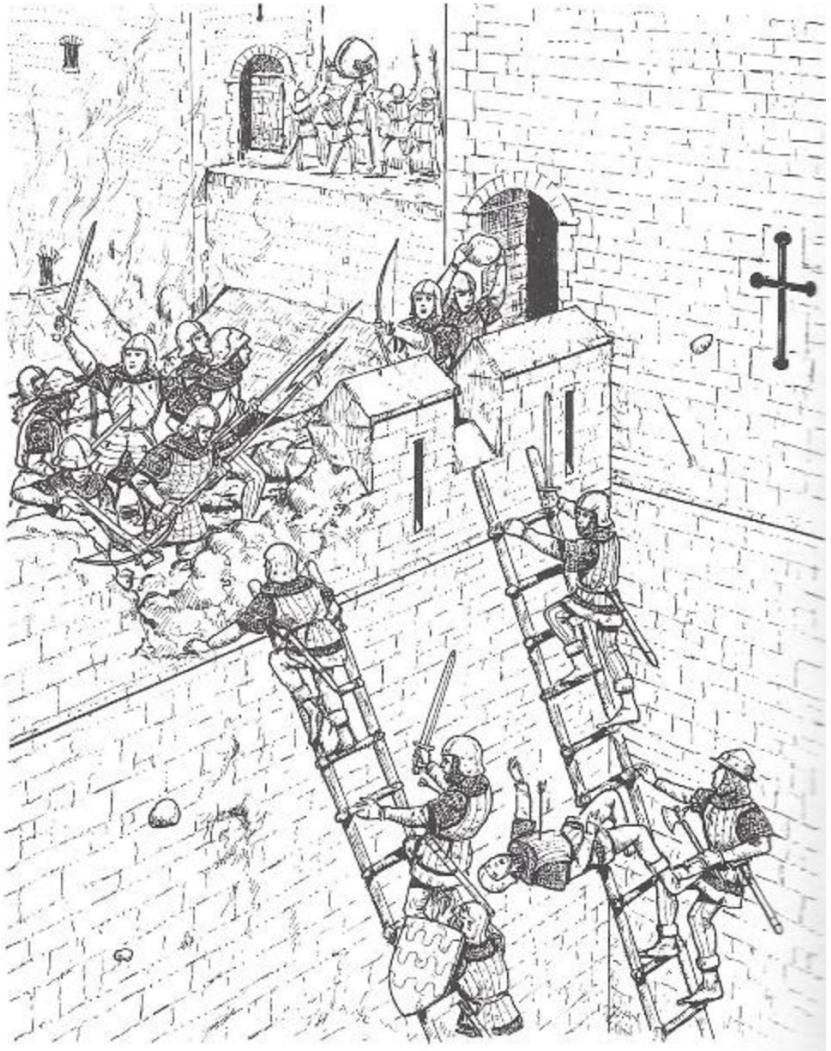
Vital Statistics vs. Mathematical Statistics

Average: 집단을 요약 → Variation: 개인(개체)들에 존재하는 차이에 관심

인구통계
평균, 비율

수리통계
분산, 관계, 추론
확률, 유의성

3 Types of Average: Mean, Median, Mode



성벽의 벽돌 갯수를 병사들이 측정한 값들

병-1	병-2	병-3	병-4	병-5	병-6	병-7	병-8	병-9
13	18	13	15	13	16	14	21	13

Q. 어떤 값을 대표값으로 선택할까?

A. 평균
(13+18+13+14+13+16+14+21+13) ÷ 9 = **15**

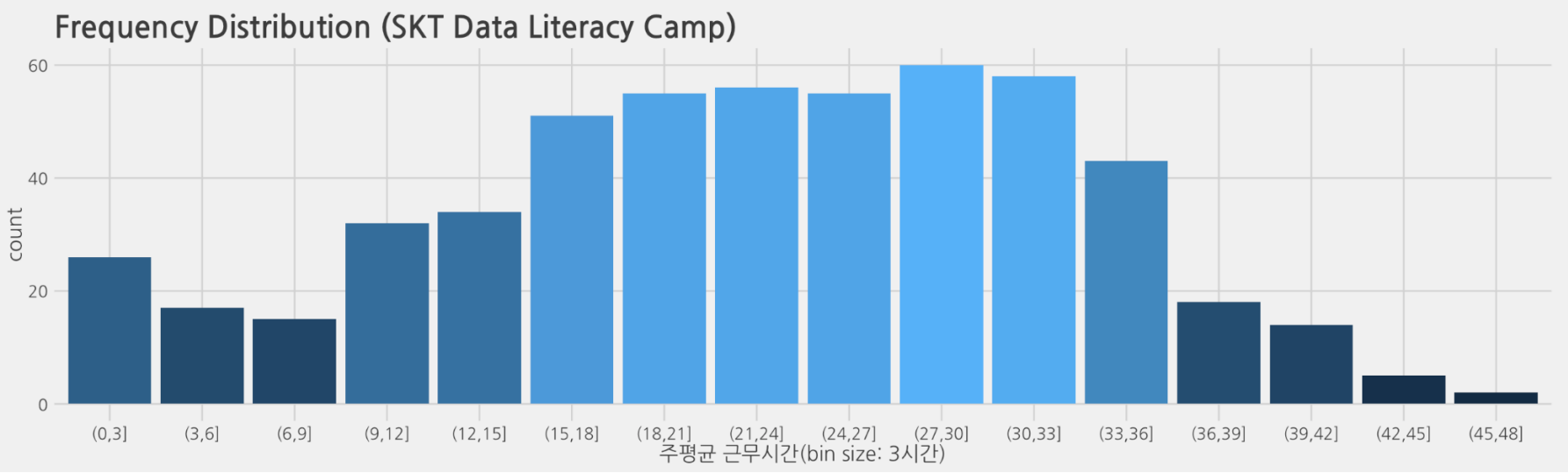
B. 중앙값
13, 13, 13, 13, **14**, 14, 16, 18, 21

C. 최빈값
13 (3번 측정; 다른 값들은 1번씩만 측정됨)

D. 선호값
16 (병-6)

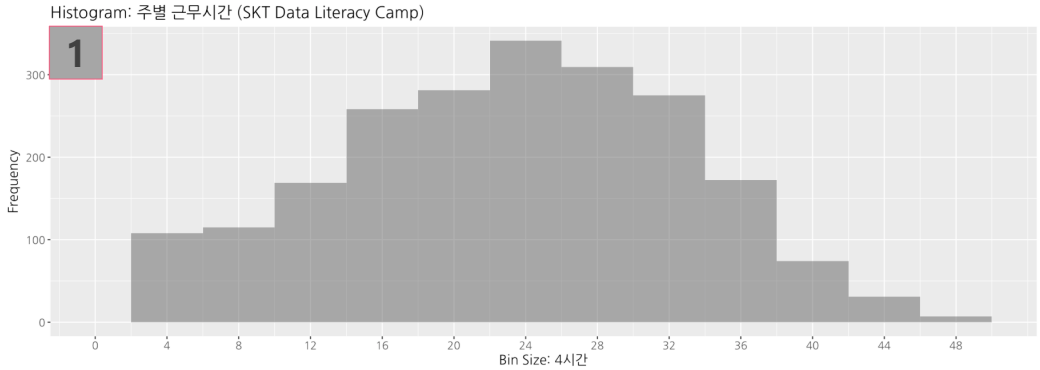
Histogram vs. Frequency Distribution Table

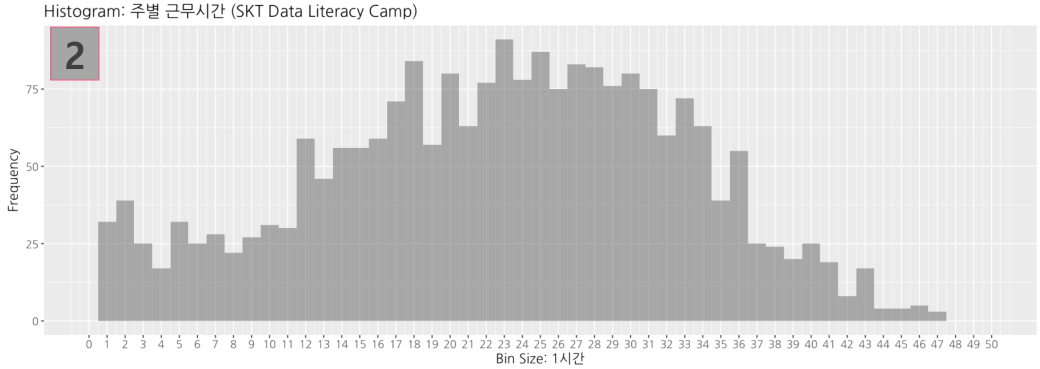
히스토그램과 도수분포표

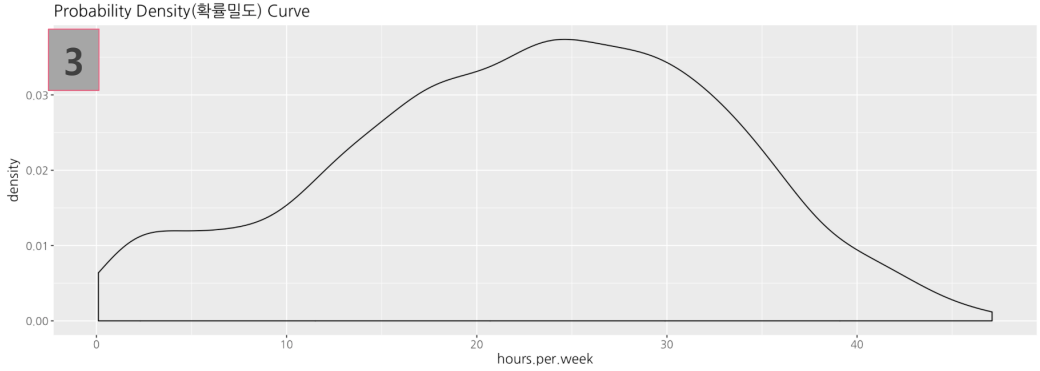


계급	(0,3]	(3,6]	(6,9]	(9,12]	(12,15]	(15,18]	(18,21]	(21,24]	(24,27]	(27,30]	(30,33]	(33,36]	(36,39]	(39,42]	(42,45]	(45,48]
빈도	26.0	17.0	15.0	32.0	34.0	51.0	55.0	56.0	55.0	**60.0**	58.0	43.0	18.0	14.0	5.0	2.0
누적 빈도	26.0	43.0	58.0	90.0	124.0	175.0	230.0	286.0	341.0	**401.0**	459.0	502.0	520.0	534.0	539.0	541.0
비율	4.8	3.1	2.8	5.9	6.3	9.4	10.2	10.4	10.2	**11.1**	10.7	7.9	3.3	2.6	0.9	0.4
누적 비율	4.8	7.9	10.7	16.6	22.9	32.3	42.5	52.9	63.1	**74.2**	84.9	92.8	96.1	98.7	99.6	100.0

Histogram vs. Density Plot







1 Histogram – Bin Size: 4시간

- 히스토그램: 도수(빈도)의 분포[도수분포표]를 차트로 표현한 것
- 계급: X축에 표현된 변수의 구간[4시간]

2 Histogram – Bin Size: 1시간

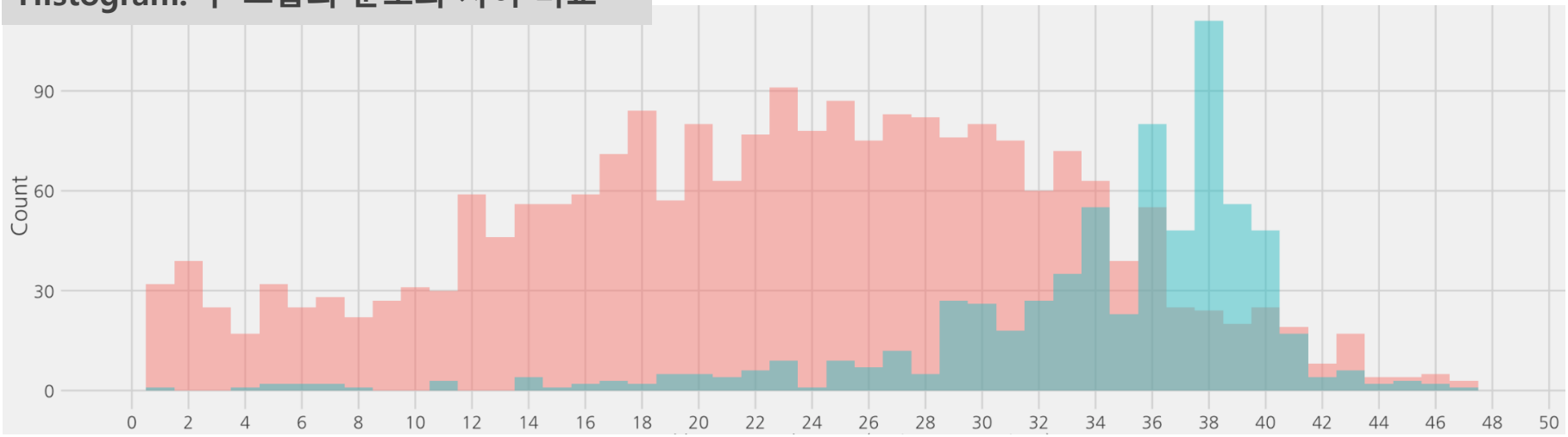
- X축 변수 구간의 크기(Bin Size)를 4시간에서 1시간으로 조정하였음

3 Probability Density Curve

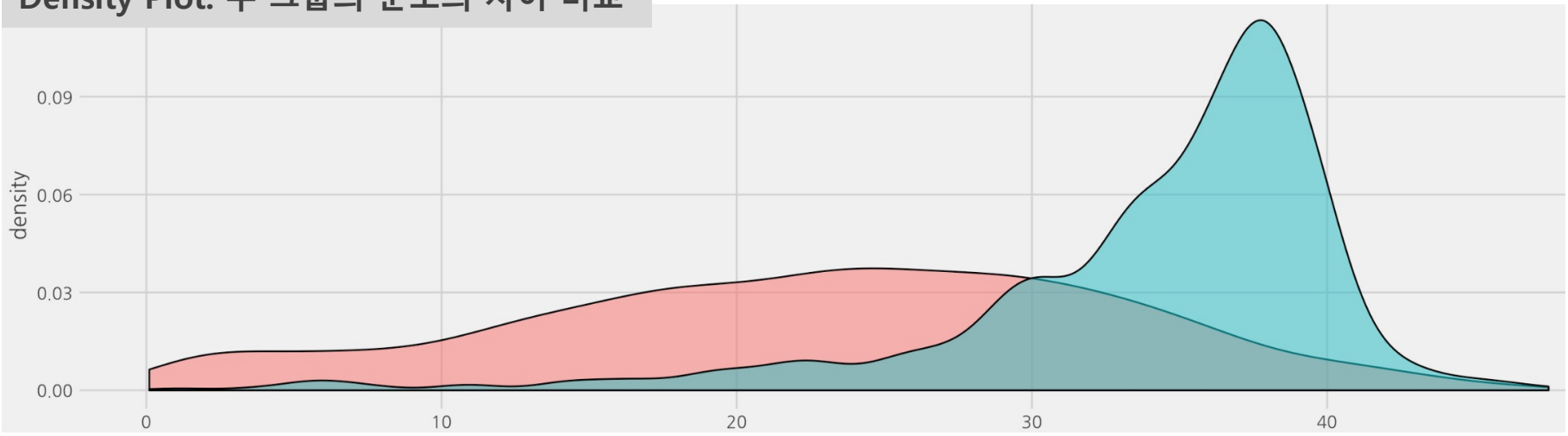
- 확률밀도: X가 연속형 변수일 경우 X값과 이에 대응하는 확률을 나타낸 그래프
- 좌측에서 X가 10~20시간 사이의 값을 가질 확률은 해당 구간의 면적과 동일함

Histogram vs. Density Plot: 서로 다른 두 집단의 분포를 비교

Histogram: 두 그룹의 분포의 차이 비교



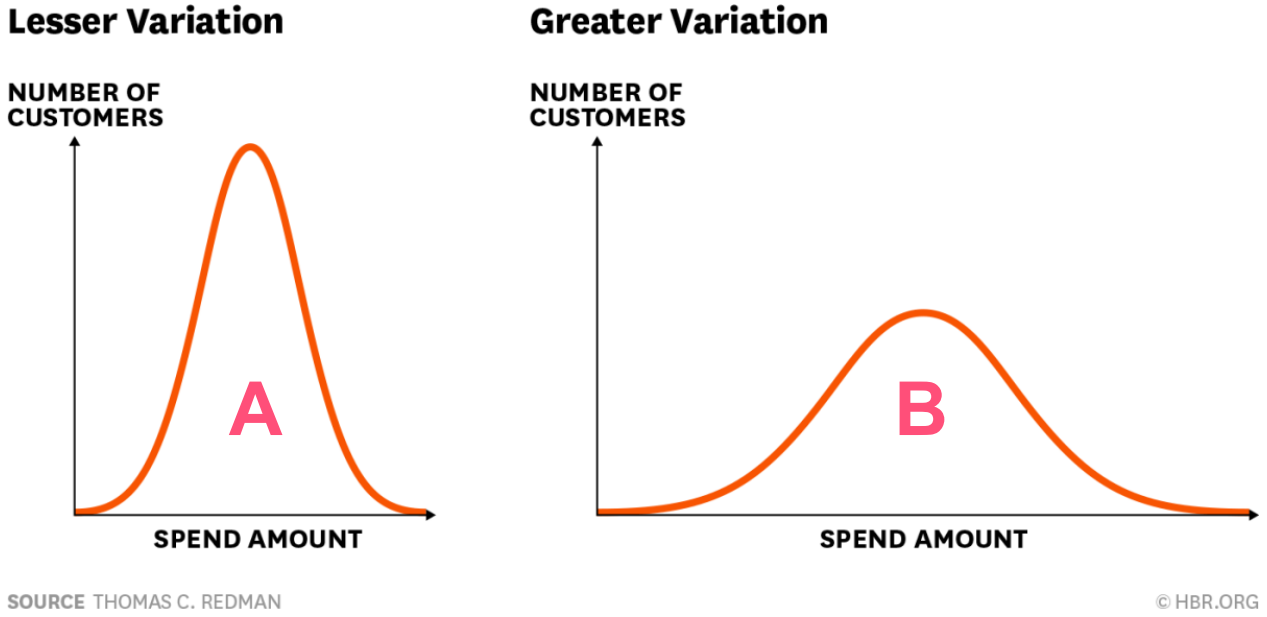
Density Plot: 두 그룹의 분포의 차이 비교



Larger Variation, Greater Sampling Error

캠페인	고객수	평균 지출
신규	30	10,000
기존	1,500	8,000

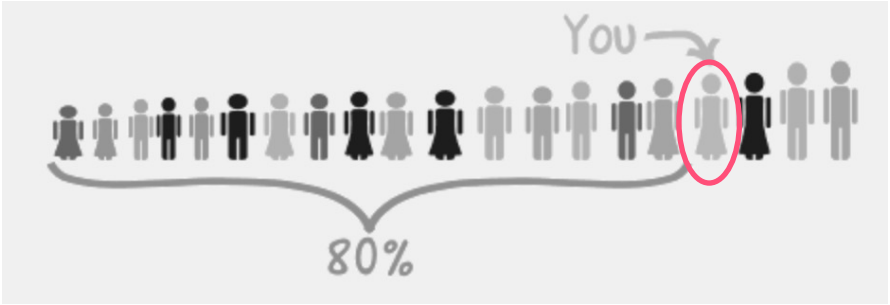
Q. 고객들(모집단)의 지출금액 분포, A와 B 중, 신규 캠페인의 효과를 주장(일반화)하기에 더 좋은 것은?



변화의 폭 ↗ 샘플 데이터 신뢰도 ↘ 평균값에 대한 확신 ↘

Percentile

Percentile: 전체 관측값들의 분포를 고려했을 때 특정값의 상대적 위치



내 키가 185cm로 20명 중
네번째로 키가 크다면

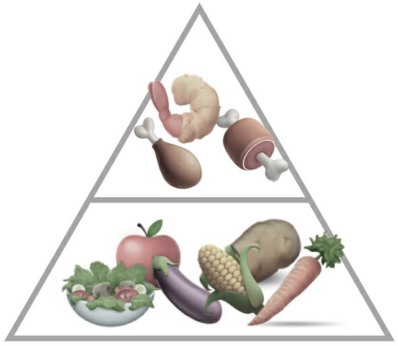
185cm = 80th Percentile

내 밑으로 80%가 있다!

Score [정렬된 점수]	Percentile Rank	Quartile [사분위]
29	8th	Q1, 1사분위 (최하위 25%)
32	17th	
38	25th	
41	33th	Q2, 2사분위 (차하위 25%)
53	42th	
54	50th	
55	58th	Q3, 3사분위 (차상위 25%)
74	67th	
93	75th	
99	83th	Q4, 4사분위 (최상위 25%)
134	**92th**	
209	100th	

Percentile 활용하여 주성분(Principal Component) 찾기

비타민, 지방, 섬유질 변수로 채소와 육류 분류하기



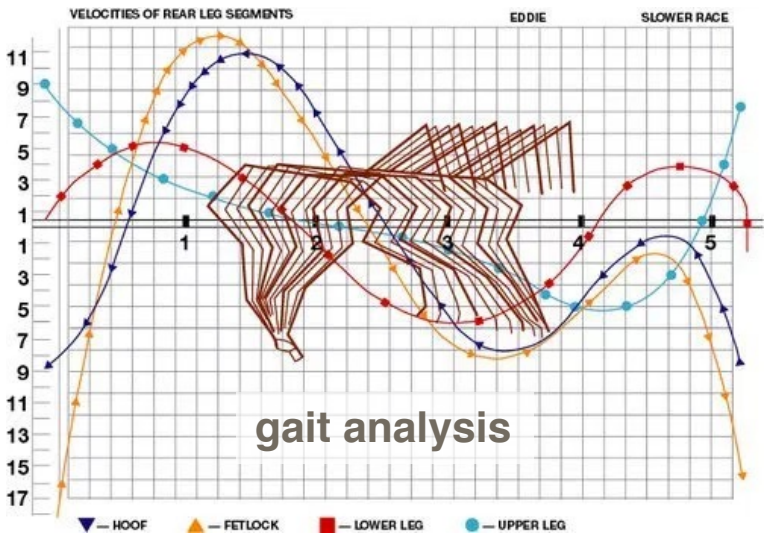
- **PCA(Principal Component Analysis):** 데이터 분류를 용이하게 하는 (=데이터가 최대한 퍼지게 하는) 주성분(Principal Component; 이 경우 Vitamin C + Fiber – Fat)을 찾는 일.
- **"Vitamin C – Fat":** Percentile 값으로 바꾸면 해당 변수를 정규화하는 효과가 있어서 서로 다른 단위를 갖는 변수들 간 연산이 가능해짐

Finding Secret Feature

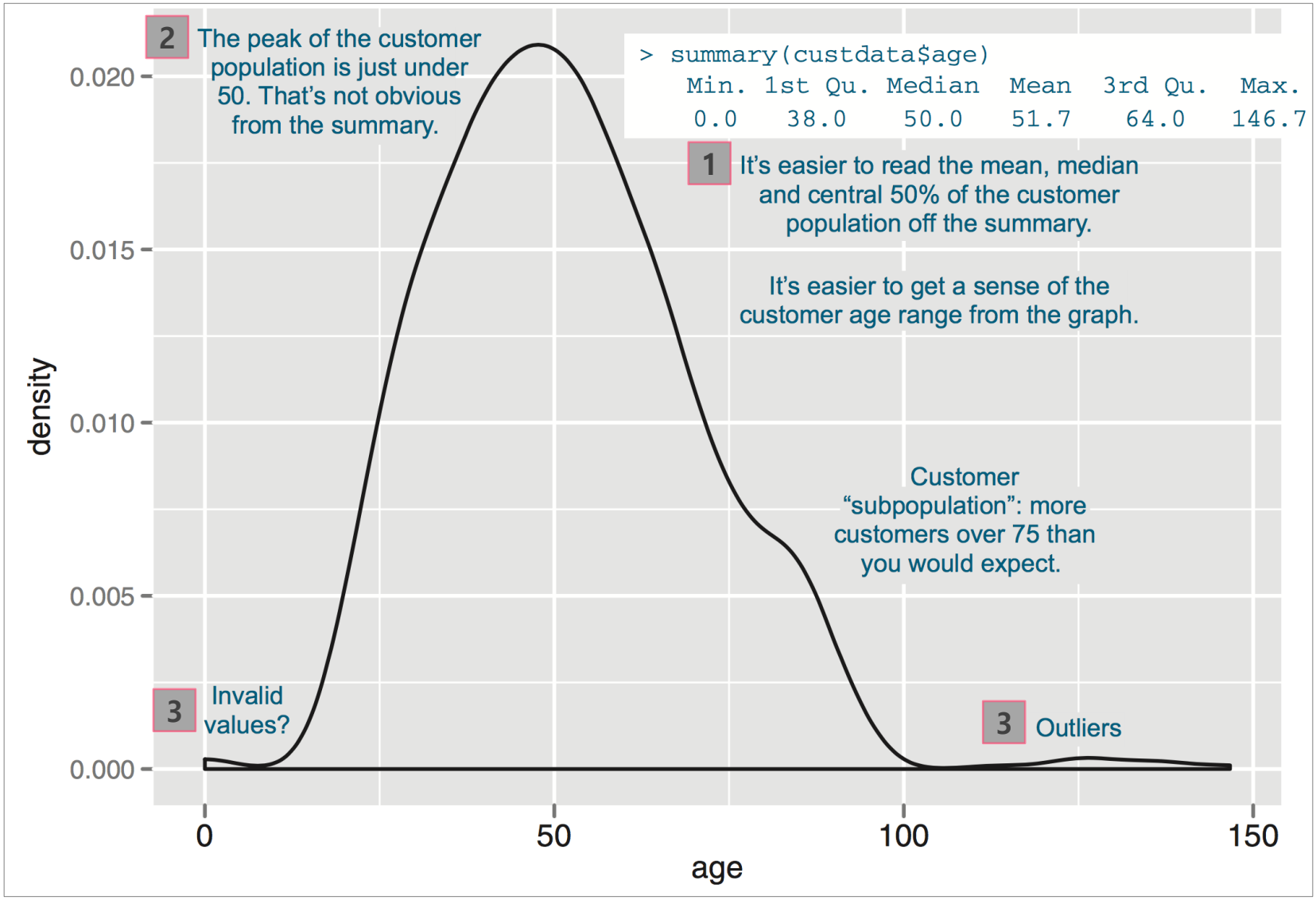
Percentile과 남들은 모르는 좋은 Feature로 돈 번 사례

American Pharoah: 2015년도에 37년 만에 Triple Crown 달성 (삼관마)
Jeff Seder: **"Sell your house. But, do NOT sell this horse."**

Variable	Percentile
Height	56%
Weight	61%
Pedigree	70%
Left Ventricle (좌심실)	**99.61%**

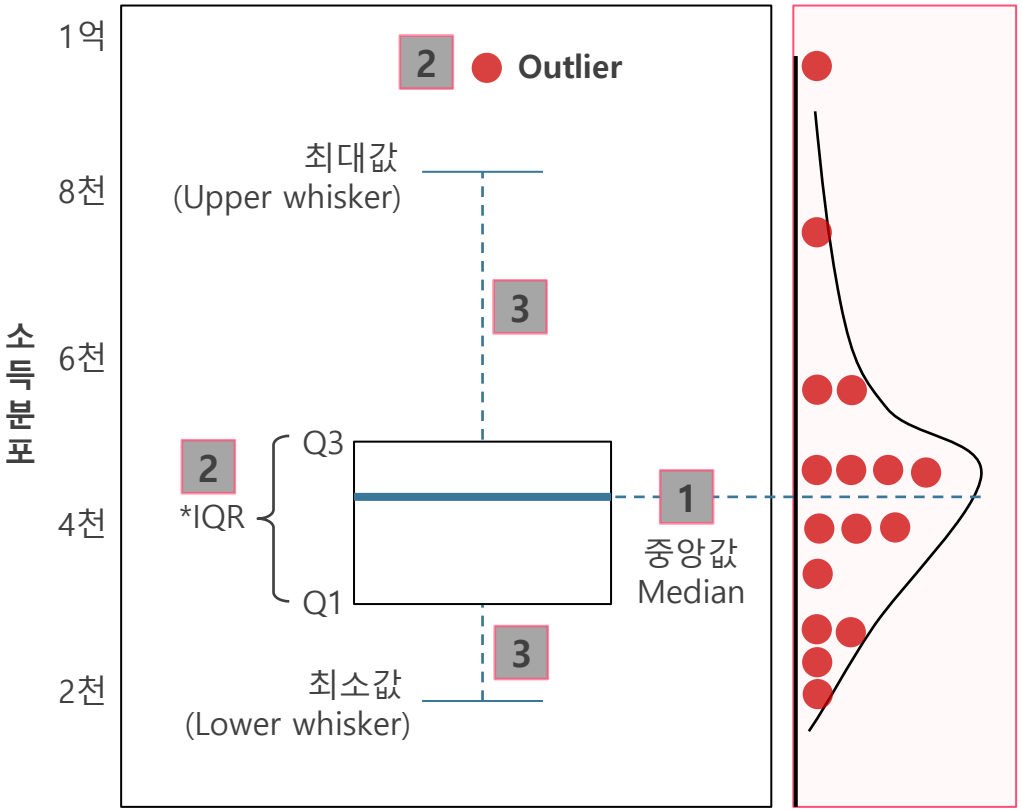


Summary Statistics and Visualization



Distribution (boxplot)

Box-and-whiskers plot (예시)



Box-and-whiskers plot 해석

1 중심 경향

- 중앙값(median) 파악

2 특이값(Outlier)

- *IQR(Inter Quartile Range)의 1.5배 이상 벗어나 있는 값들을 Outlier로 정의

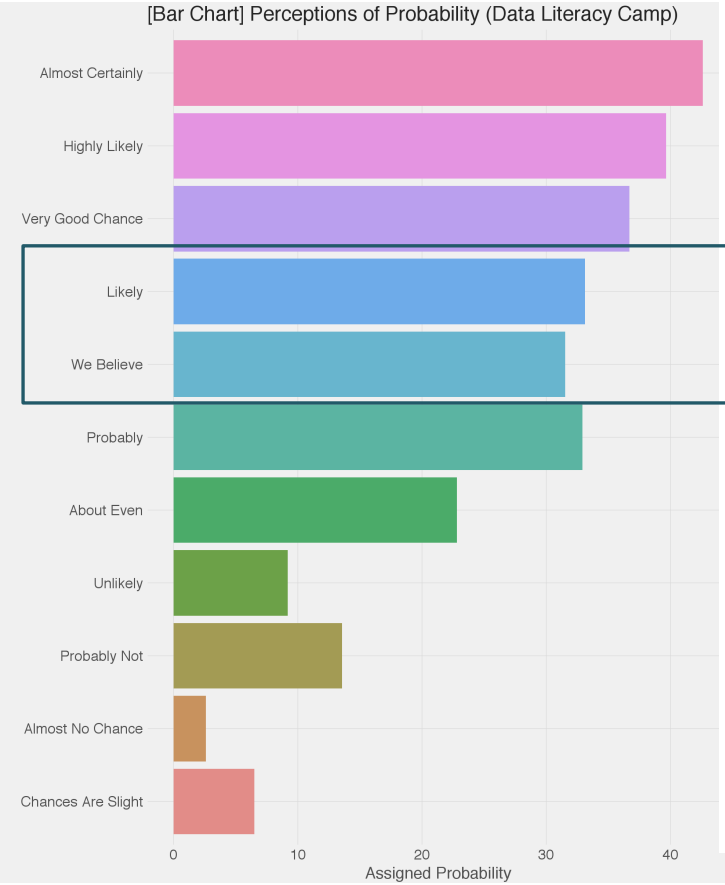
3 대칭성 및 분포

- 대칭성(symmetry): 최대값과 최소값까지의 수염 길이 비교

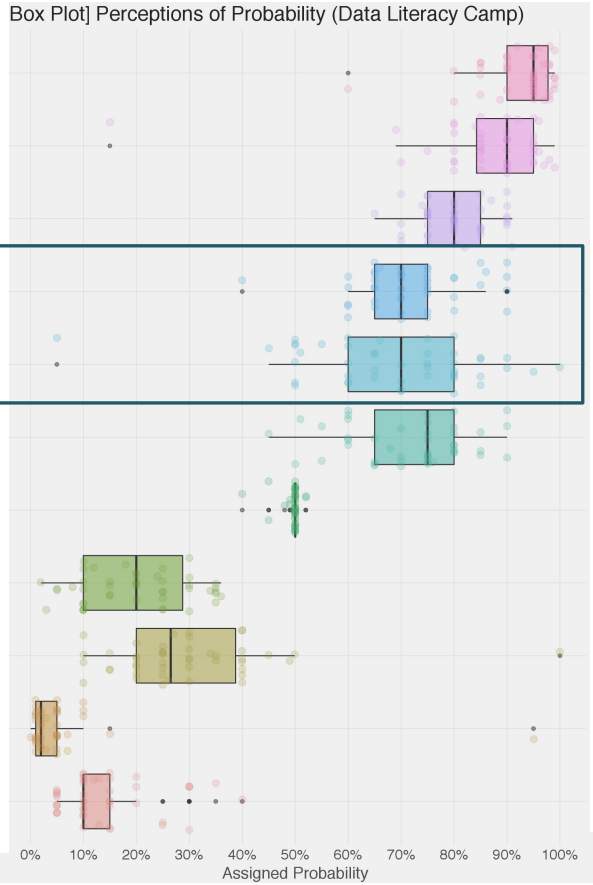
Data Description

- 평균: 요약값을 사용하여 서로 다른 범주들의 순위를 비교
- 분포: 퍼진 정도를 관찰하여, 개별 범주 내 차이를 이해

Bar Chart (평균)



Box Plot (분포)



Density Plot (분포)